Ahmedabad University

MAT502 Advanced Statistics

Project Report

Analysis to find the relationship between consumption of beverages and the sleeping routine of the individuals

Submitted to faculty: Prof. Shashi Prabh

Date of Submission: 18th April 2023

Student Details

Roll No.	Name of the Student	Name of the Program
AU2040004	Priya Jani	BTech CSE
AU2040025	Yagnesh Patel	BTech CSE

2022-2023 (Winter Semester)

Motivation

We need a good night's sleep for our physical and emotional health. Yet, how we live our lives, especially the beverages we choose to drink, can significantly impact how well and how long we sleep. Humans consume a variety of liquids, including caffeinated beverages like soda, coffee, and tea. While some people think certain drinks improve their sleep, others believe they disturb their sleep cycles. If there is a connection between them, we want to know about it.

We can learn how certain drinks affect our sleep by understanding the relationship between drinking habits and sleeping patterns, which is an exciting field of research. By looking at this correlation, we may find out which drinks are most likely to disrupt our sleep patterns and which are most helpful for getting a good night's rest.

Objective

Our main objective is to find out the correlation between beverages and the sleep patterns of individuals.

Some other objectives are to check whether any column of the dataset behaves randomly or not. Also, apart from narrowing the relationship between beverages and sleep, we would like to check if some other relations come into play while working on the process.

Methodology-Sampling Methods and Experiment Design-Hypothesis, Data Collection

We have used convenience sampling for our case study. Convenience sampling is a non-probability sampling strategy that involves choosing participants based on their accessibility, availability, and desires to participate in a study, making data collection easy.

Data Used:

We have collected and gathered information from people of different categories based on age and gender. We have collected data from more than 60 people. The data mainly collected from them was based on the number of different beverages they consume, their daily quantities, and their sleeping routines.

Data collection method:

We asked different questions from different people. We have approached people randomly as well. Since we want to determine whether there is a relationship between beverage consumption and sleep, the following questions were asked the people:

1. How often do you drink these beverages(Tea, Coffee, Energy Drinks, Soft Drinks, Milk, Juice) daily? (where numbers, i.e., "0" to "5 or more," represent the number of cups/cans/glasses)
2. Regularly, how does your sleeping schedule change when you consume TEA?
3. Regularly, how does your sleeping schedule change when you consume COFFEE?
4. On the regular basis, how does your sleeping schedule change when you consume ENERGY DRINKS?
5. Regularly, how does your sleeping schedule change when you consume SOFT DRINKS?
6. Regularly, how does your sleeping schedule change when you consume MILK?

7. Regularly, how does your sleeping schedule change when you consume JUICE (any)?
8. How well is your sleeping schedule?
9. Do you think your sleeping routine is affected by the beverages you consume?
10. How is your sleeping schedule when you don't consume beverages that you prefer the most?

Possibilities of response bias:

- Since we didn't know much about the routines and habits of people, we blindly considered their responses. There might be a chance that people would have given us the wrong information.
- Some people may have given the wrong information about the number of cups they consume different beverages regularly.

Possibilities of non-response bias:

- On average, approaching people and asking them questions, we found that few people (8-9) just ignored it and gave excuses.
- One of the reasons could be that they might have thought of information getting leaked or been busy when questions were asked.

Data Preprocessing and Encoding:

As most of the data was categorical, we did encoding to convert it into numerical. The following is the outcome.

```
data.to_csv("DataEncoding.csv")
data = data.iloc[:,1:]
```

```
8]  data.head()
```

nergyDrinks_sleepingPattern	SoftDrinks_sleepingPattern	Milk_sleepingPattern	Juices_sleepingPattern	SleepingSchedule	BeveragesAfftectingSleep
2	2	2	2	2	1
3	2	2	2	2	1
2	2	2	2	1	0
0	0	2	2	4	0
2	2	1	2	2	0

Fig1: Data Encoding

Data Visualization and Analysis

From the responses of 72 people, we did the descriptive analysis and visualization of data based on different questions.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 72 entries, 0 to 71
Data columns (total 16 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Name                         72 non-null     object
 1   Age                          72 non-null     object
 2   TEA_consumption              72 non-null     int64
 3   Coffee_consumption           72 non-null     int64
 4   EnergyDrinks_consumption     72 non-null     int64
 5   SoftDrinks_consumption       72 non-null     int64
 6   Milk_consumption             72 non-null     int64
 7   Juices_consumption           72 non-null     int64
 8   TEA_sleepingPattern          72 non-null     object
 9   Coffee_sleepingPattern       72 non-null     object
 10  EnergyDrinks_sleepingPattern 72 non-null     object
 11  SoftDrinks_sleepingPattern   72 non-null     object
 12  Milk_sleepingPattern         72 non-null     object
 13  Juices_sleepingPattern       72 non-null     object
 14  SleepingSchedule             72 non-null     object
 15  BeveragesAfftectingSleep     72 non-null     object
dtypes: int64(6), object(10)
memory usage: 9.1+ KB
```

Fig2: Descriptive Statistics of the Columns

The age composition of the people was as follows:

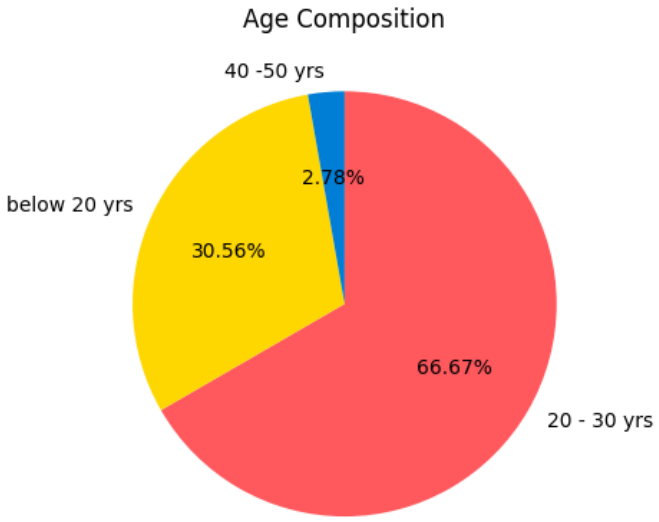

Fig3: Age Composition

Mostly our responses were from people of age 20-30 years. It could be considered as a bias as it can temper our results.

The below figure (Fig4) shows the number of people and the quantitative measure of cups/cans/glasses they consume on regular basis.

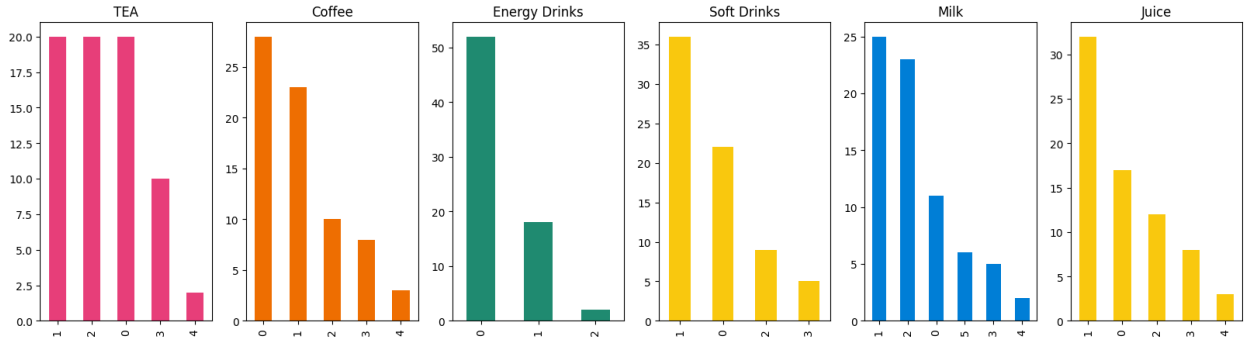

Fig4: People opting the number of cups/cans/glasses of a particular beverages

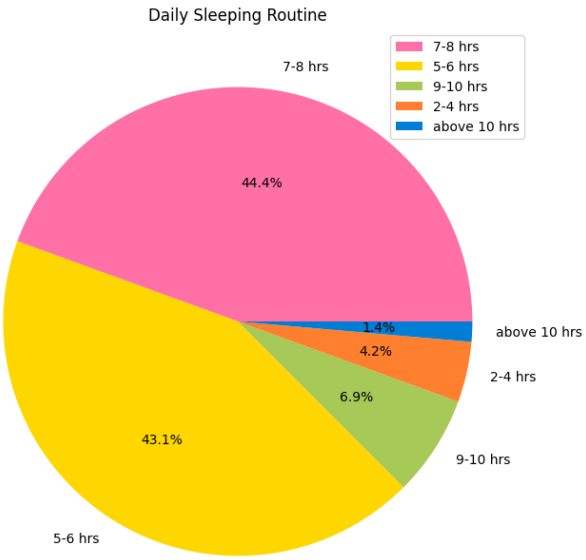

Fig5: Sleeping Routine Composition

The above figure (Fig5) shows the number of people and their sleeping routines on the basis of hours.

The below figures (Fig6 to Fig11) show the sleeping pattern with respect to beverages in particular.

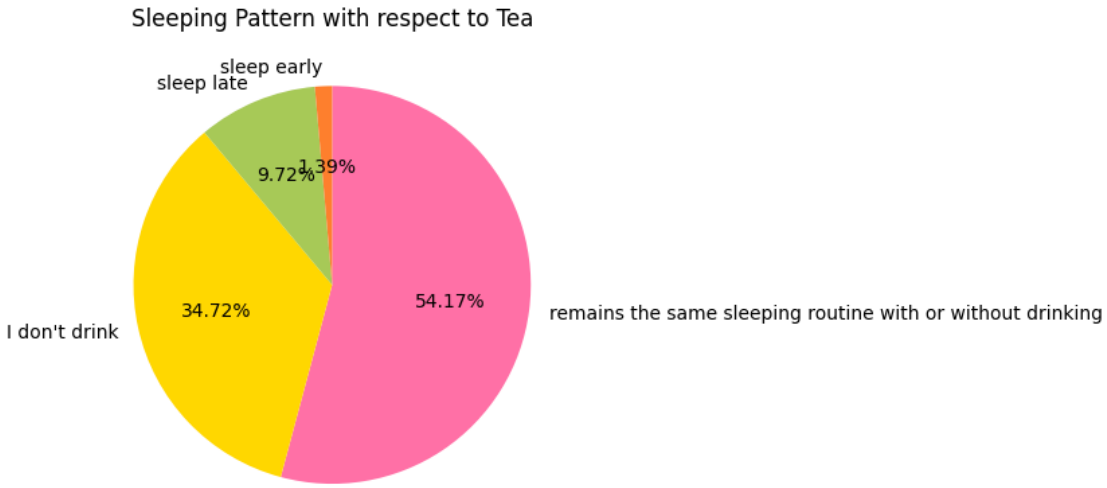

Fig6: Sleep Pattern with respect to Tea

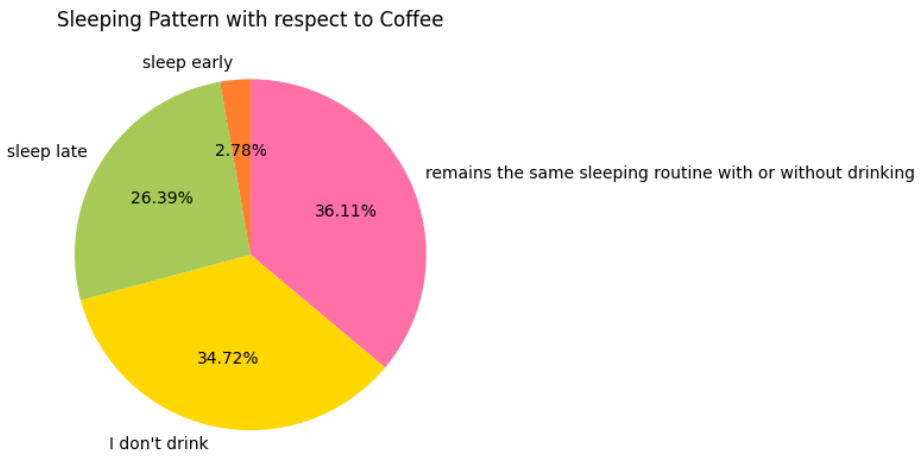

Fig7: Sleep Pattern with respect to Coffee

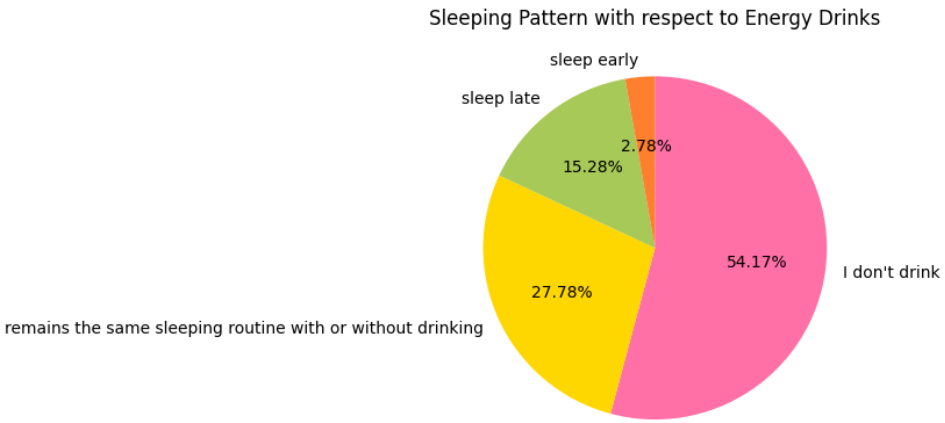

Fig8: Sleep Pattern with respect to Energy Drinks

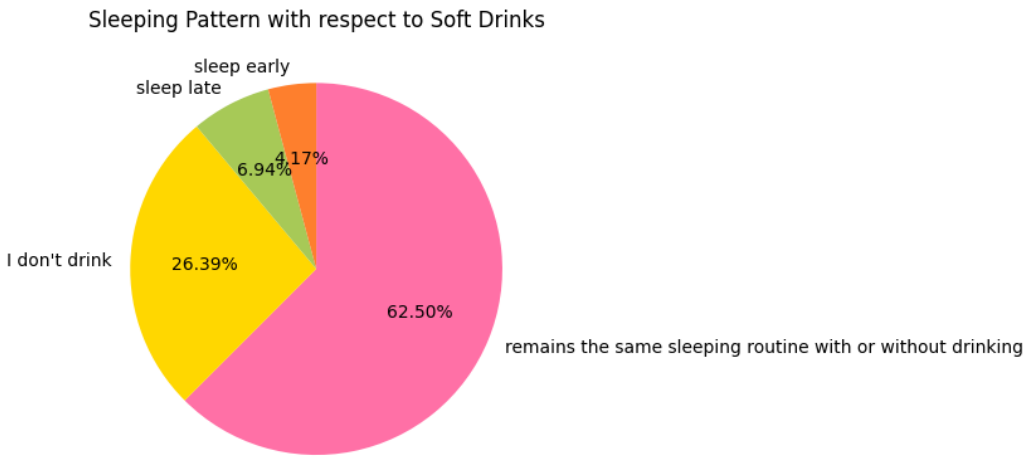

Fig9: Sleep Pattern with respect to Soft Drinks

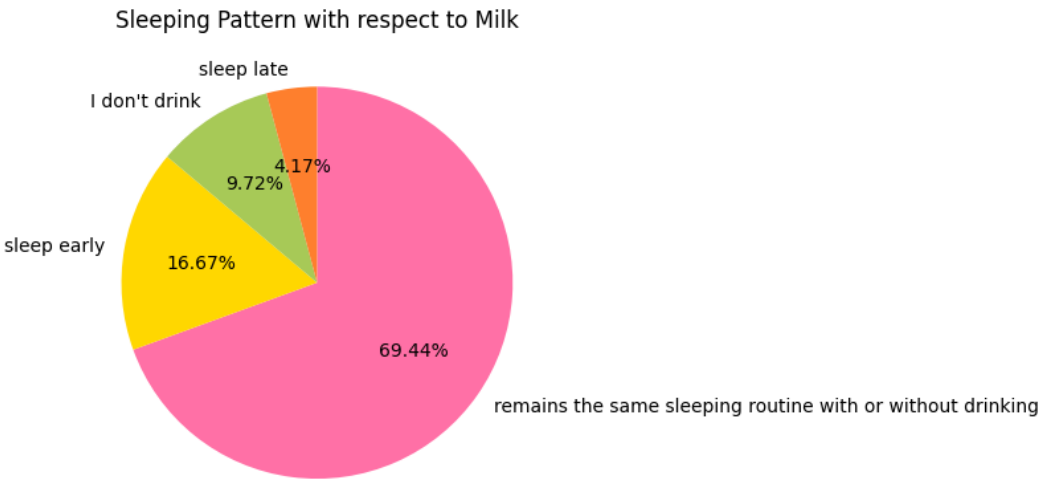

Fig10: Sleep Pattern with respect to Milk

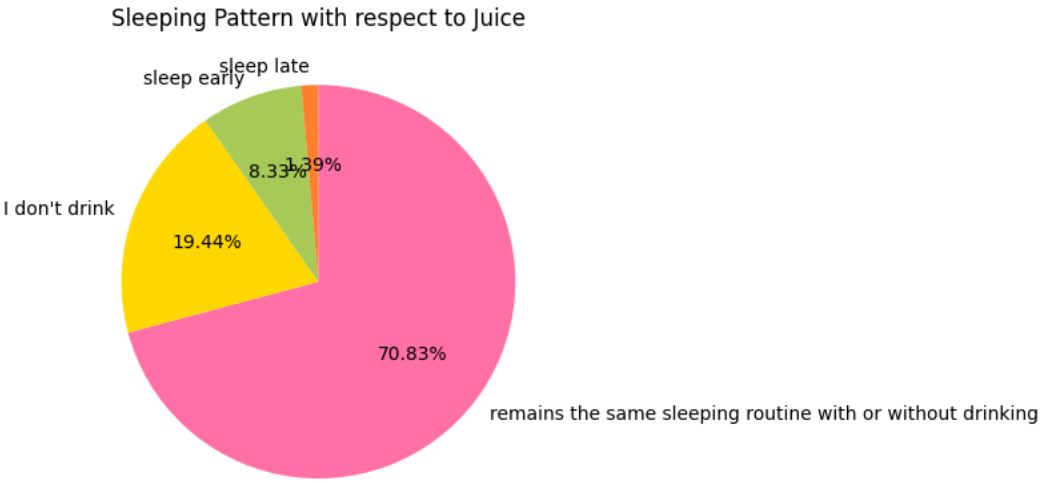

Fig11: Sleep Pattern with respect to Juice

Hypothesis

We used Python to perform a chi-square test of independence to identify the relationship between the number of different beverages consumed and sleeping patterns.

The chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables. It is based on the chi-square distribution, which is a probability distribution that describes the relationship between the observed frequencies of an event and the expected frequencies of that event.

- Test of independence:
  - $\chi^2 = \sum [(O_ij - E_ij)^2 / E_ij]$
  -
  - Where:
  -
  - χ^2 is the chi-square test statistic
  - O_ij is the observed frequency for the i-th row and j-th column
  - E_ij is the expected frequency for the i-th row and j-th column, calculated as (total count of row i * total count of column j) / total sample size
  - $\sum$ is the sum of all rows and columns being tested

Hypothesis 1

NULL Hypothesis (H0): There is no significant association between tea consumption and an individual's sleeping patterns.

Alternative Hypothesis (H1): Tea consumption and an individual's sleeping patterns are significantly associated.

```
contingency_table = pd.crosstab(data['TEA_consumption'], data['TEA_sleepingPattern'])

# perform the chi-squared test
stat, p, dof, expected = chi2_contingency(contingency_table)

# print the results
print('Chi-Squared Statistic: ', stat)
print('p-value: ', p)
print('Degrees of Freedom: ', dof)
# print('Expected
```

```
Chi-Squared Statistic:  66.88351648351647
p-value:  1.2205201447035471e-09
Degrees of Freedom:  12
```

Fig12: Chi-Square Test for Tea and the corresponding Sleeping routine

The results:

- Chi-square statistic: 66.88
- p-value: 1.221e-09
- Degrees of freedom = 12

As the p-value<0.01 is highly significant, the NULL hypothesis is rejected, meaning there is a significant association between tea consumption and an individual's sleeping patterns for the data we collected.

Similarly, we did it for all the beverages and found that beverage consumption affects the sleeping routine of an individual.

Similarly, the other hypotheses were made for different beverages in particular and the following are the results:

```
[9]  contingency_table = pd.crosstab(data['Coffee_consumption'], data['Coffee_sleepingPattern'])

     # perform the chi-squared test
     stat, p, dof, expected = chi2_contingency(contingency_table)

     # print the results
     print('Chi-Squared Statistic: ', stat)
     print('p-value: ', p)
     print('Degrees of Freedom: ', dof)
     # print('Expected

     Chi-Squared Statistic:  57.30749264465512
     p-value:  6.955344980511283e-08
     Degrees of Freedom:  12
```

Fig13: Chi-Square Test for Coffee and the corresponding Sleeping routine

```
contingency_table = pd.crosstab(data['EnergyDrinks_consumption'], data['EnergyDrinks_sleepingPattern'])

# perform the chi-squared test
stat, p, dof, expected = chi2_contingency(contingency_table)

# print the results
print('Chi-Squared Statistic: ', stat)
print('p-value: ', p)
print('Degrees of Freedom: ', dof)
# print('Expected

Chi-Squared Statistic:  45.67582929890622
p-value:  3.4347079464854817e-08
Degrees of Freedom:  6
```

Fig14: Chi-Square Test for Energy Drinks and the corresponding Sleeping routine

```
contingency_table = pd.crosstab(data['SoftDrinks_consumption'], data['SoftDrinks_sleepingPattern'])

# perform the chi-squared test
stat, p, dof, expected = chi2_contingency(contingency_table)

# print the results
print('Chi-Squared Statistic: ', stat)
print('p-value: ', p)
print('Degrees of Freedom: ', dof)
# print('Expected

Chi-Squared Statistic:  51.24984582668793
p-value:  6.264358113742697e-08
Degrees of Freedom:  9
```

Fig15: Chi-Square Test for Soft Drinks and the corresponding Sleeping routine

```
contingency_table = pd.crosstab(data['Milk_consumption'], data['Milk_sleepingPattern'])

# perform the chi-squared test
stat, p, dof, expected = chi2_contingency(contingency_table)

# print the results
print('Chi-Squared Statistic: ', stat)
print('p-value: ', p)
print('Degrees of Freedom: ', dof)
# print('Expected
```

```
Chi-Squared Statistic:  43.03096555618295
p-value:  0.0001557179143996847
Degrees of Freedom:  15
```

Fig16: Chi-Square Test for Milk and the corresponding Sleeping routine

```
contingency_table = pd.crosstab(data['Juices_consumption'], data['Juices_sleepingPattern'])

# perform the chi-squared test
stat, p, dof, expected = chi2_contingency(contingency_table)

# print the results
print('Chi-Squared Statistic: ', stat)
print('p-value: ', p)
print('Degrees of Freedom: ', dof)
# print('Expected
```

```
Chi-Squared Statistic:  31.636678200692046
p-value:  0.0015740339313950063
Degrees of Freedom:  12
```

Fig17: Chi-Square Test for Juices and the corresponding Sleeping routine

For all the above tests (Fig12 - Fig17), we can see that p-value is lower than 0.01 for all the tests, highly significant, meaning there is a significant association between beverage consumption and an individual's sleeping patterns for the data we collected.

Findings/Results

- For all the Chi-Square Hypothesis tests (Fig12 - Fig17), we can see that p-value is lower than 0.01 for all the tests, highly significant, meaning there is a significant association between beverage consumption and an individual's sleeping patterns for the data we collected.
- To find the correlation between the beverages and sleeping routine, we have made a correlation matrix to cross-check the hypothesis. The below figure (Fig18) shows the correlation between the features:

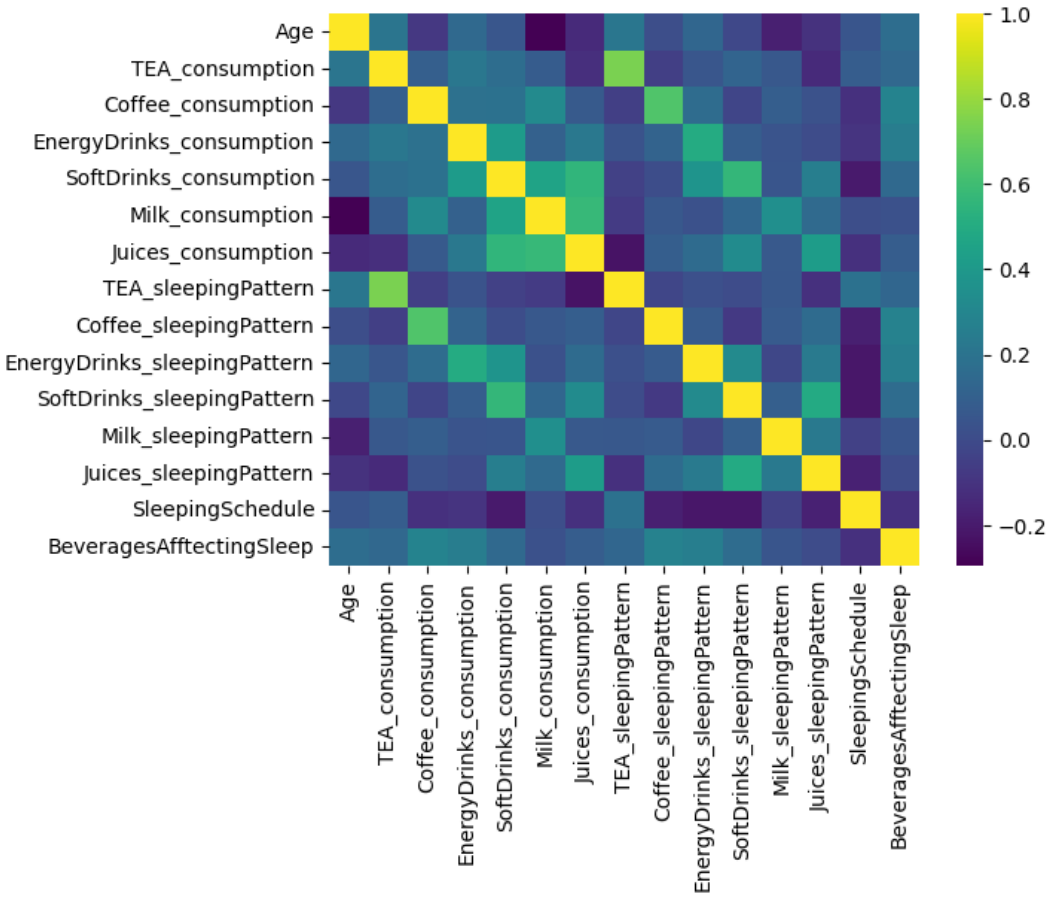

Fig18: Correlation Matrix

- We can see a high correlation between tea consumption and tea sleeping pattern. For other beverages, the value of the correlation decreases, but overall, it can be concluded that beverage consumption affects the sleeping routine of an individual.

Conclusion

- From the first hypothesis test, we can conclude that tea consumption affects the sleeping routine of an individual.
- From the second hypothesis test, we can conclude that coffee consumption affects the sleeping routine of an individual.
- Hence performing a chi-square test of independence on all the beverages, we can conclude that overall beverage consumption affects the sleeping routine of an individual.
- From the correlation matrix also, we can conclude that beverage consumption affects the sleeping routine of an individual.
- Also, there could be a possible bias in the results as age could be one of the factors affecting the results. As we had more responses from the people of the age group 20-30 years, it can be the case that it may overshadow the other age groups and lead to different results.

Improvements to get better results

- An Equal number of responses to be taken from people of different age groups to remove biases.
- Other parameters/information could have been collected to get more inferences and different observations.
- Other tests could have been applied to get more vision for the same topic.

The files for the above project are available on this GitHub link:

https://github.com/yjp1406/advStats-project